Subtle differences in language experience moderate performance on language-based cognitive tests

Maury Courtland[†], Aida Davani[‡], Melissa Reyes[*], Leigh Yeh[‡], Jun Leung[*], Brendan Kennedy[‡], Morteza Dehghani[*‡], and Jason Zevin[*†]

[†] Department of Linguistics
[‡] Department of Computer Science
[*] Department of Psychology
University of Southern California
{landerpo, mostafaz, reyesmel, leighyeh, junyenle, btkenned, mdehghan, zevin}@usc.edu

Abstract

Cognitive tests used to measure individual differences are generally designed with *equality* in mind: the same "broadly acceptable" items are used for all participants. This has unknown consequences for *equity*, particularly when a single set of linguistic stimuli are used for a diverse population of language users. We hypothesized that differences in language variety would result in disparities in psycholinguistically meaningful properties of test items in two widely-used cognitive tasks, resulting in large differences in performance. As a proxy for individuals' language use, we administered a self-report survey of media consumption. We identified two substantial clusters from the survey data, roughly orthogonal to *a priori* groups recruited into the study (university students and members of the surrounding community). We found effects of both population and cluster membership. Comparing item-wise differences between the clusters' language models did not identify specific items driving performance differences.

Introduction

Cognitive tests are increasingly used in research on individual differences. For example, a number of recent studies reported correlations between speech perception in noise and working memory (for meta-analysis, see: Dryden, Allen, Henshaw, and Heinrich 2017). Widely used tests for both (Daneman & Carpenter, 1980; Kalikow, Stevens, & Elliott, 1977) were developed without much regard for potential individual differences in language experience, however. This raises the possibility that at least some of the variability in these tasks is related to differences in participants' language experience, as demonstrated in studies of higher-level language processing (Moore & Gordon, 2015; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). Currently, it remains unclear how much this robust correlation between the two tasks – found in 26 of the 30 studies surveyed by Akeroyd (2008) – reveals a correlation between the target constructs or a latent variable of language experience.

Linguists have long considered the communicative capacities of every language to be equal and equally expressive (Joseph & Newmeyer, 2012; Pellegrino, Coup, & Marsico, 2011). Guidelines from the American Speech-Language-Hearing association on cultural competence encourage clinicians to take cultural variables into account in assessing and treating language disorders and differences (American Speech-Language-Hearing Association (ASHA), n.d.). Despite these commitments in allied fields, and the demonstrable existence of multiple American Englishes (e.g. see for review: Labov, Ash, and Boberg 2006; Schneider and Kortmann 2004), most cognitive tests assume "Mainstream" American English (MAE) as a default in the construction of stimuli, potentially confounding cognitive test performance with experience and fluency in MAE. Conversely, language experience is not deterministically related to the usual features that define distinct "dialects" – region, ethnicity, class, etc. People are cosmopolitan and idiosyncratic in the language experiences they seek out, and as a result, may be familiar with multiple language varieties, with potential consequences for their performance on cognitive tests.

Statistical learning, hypothesized to underlie much of language development (Elman, 2001; Seidenberg & MacDonald, 1999), is driven by patterns in language input. Given different input, then, language learners will necessarily construct different distributional models to generate and process speech and language. Online speech and language processing relies heavily on learned statistical regularities to facilitate top-down anticipatory processes. This is evidenced by the effects of surprisal observed when these anticipations are violated (Federmeier, Mai, & Kutas, 2005; Kutas & Hillyard, 1980, 1984). Given the highly demanding nature of online speech and language processing, anticipatory mechanisms help lessen the cognitive effort needed to accomplish the task. The greater the difference between the listener or reader's language model and the statistics of the language material they are processing, the greater the cognitive burden on the listener. For example, intelligibility levels in noise are better for one's own dialect than for a familiar, but less commonly encountered dialect (Clopper & Bradlow, 2008). In children who prefer a non-mainstream English, familiarity with "school English" is associated with performance on literacy tests (Charity, Scarborough, & Griffin, 2004).

The current research examines the effect of variability in language experience on cognitive tests. We hypothesized that measuring people's language experience indirectly, by having them complete a "media diet" survey, would allow us to identify distinct clusters of individuals based on their viewing, listening, and reading habits. We expect these clusters to only loosely covary with the demographic factors that commonly define distinct "dialect" groups. This new measure

of language differences between participants thus provides a novel aspect of individual variability that we expect to moderate performance on language-based cognitive tasks. As this measure probes the role of language directly, it may be more informative in predicting task performance variability than standard demographic information. To test this we recruit from two populations that differ along traditional demographic lines: USC undergraduates – typically high-SES students pursuing higher education (*Facts and Figures | About USC*, n.d.) – and members of the downtown Los Angeles community – mostly African American and Latinx lower-SES individuals, many of whom not pursuing education beyond high school (e.g. the zip code 90062: US Census Bureau n.d.). We administer the aforementioned functional hearing and working memory tasks and expect survey responses to at least partly predict variability in task performance. As we expect this effect to be linguistic, we also predict that language models trained on the media sources will predict participants' behavioral performances.

Methods

Participants

We recruited participants from the USC undergraduate population (N=70) and on a local community college campus (Los Angeles Trade-Technical College, N=25). USC students participated in exchange for course credit and community participants were compensated for their time at $15 per hour, prorated at 20 minute intervals. No requirements were placed on age, but due to recruitment populations, 80% of participants were between the ages of 19 and 26 (mean=22, std=6.25).

Cognitive Tests

To test participants' language abilities, we used the reading span task (Daneman & Carpenter, 1980) that was developed to assess verbal working memory and the speech perception in noise task (SPiN, Kalikow et al. 1977) that was developed to assess functional hearing. The reading span task presents sets of sentences to be read aloud while participants maintain the last word of each sentence in memory. At the end of a set, participants are tested on how many sentence-final words they can recall, and set length is increased until they cannot complete the task. Testing is terminated when participants cannot completely recall any of the three sets of sentences at a particular set length. The SPiN consists of short sentences presented over headphones in 12 talker babble (6 female, 6 male). Participants must identify the final word of the target sentence. We used recordings from the Nationwide Speech Project (Clopper & Pisoni, 2006) to create the stimuli and present trials at +6dB SNR which produced large individual differences in accuracy in pilot results. We choose these tests due to their importance as widely used individual difference measures in clinical populations to diagnose age-related decline (Byrne, 1998), aphasia (Caspari, Parkinson, LaPointe, & Katz, 1998), Alzheimer's (Kempler, Almor, Tyler, Andersen, & MacDonald, 1998), and schizophrenia (Stone, Gabrieli, Stebbins, & Sullivan, 1998). We also choose these tests for the important – but often unacknowledged – role language processing is likely to play in both.

Survey

We constructed an online survey (approx. 20 minutes long) that probes participants' current and formative media consumption habits, elicits short language production passages, and collects basic demographic information. We use this tool to glean each participant's media diet, which forms the basis for later linguistic grouping and analysis. We use the language obtained from the sources participants report as a model for participants' actual language input and a proxy of language experience.

Equipment

Subjects sat in a noise attenuating booth and participated in the survey and behavioral tasks on a desktop PC computer. USC participants were allowed to complete the survey online prior to their lab session. Participants first completed the reading span task, followed by the SPiN, and finally the survey. The reading span task was administered and scored by a researcher to ensure subjects read aloud continuously. Upon completion of each sentence, the researcher advanced the display to the next sentence in the set and solicited verbal responses at the end of each set. After a brief training phase, participants were not given feedback on their performance and were not told their failure had caused the end of the test, simply that it had ended. The SPiN test was administered using Paradigm experiment software; participants typed their responses into a free-response text box. Trials began after a 500ms delay once participants had submitted their response. Stimuli were presented at a comfortable level, standard across participants.

Clustering

We create a media source space in which each dimension represents a reported source (e.g. movie) collected in our survey. Each participant is thus represented as a binary vector in this space, with 1s in dimensions corresponding to sources they consume, and 0s in those they do not. To ensure each dimension is informative (and reduce the dimensionality), we only represent sources reported 10 or more times – thus avoiding dimensions that would only differentiate a few participants (i.e. the rest would all receive 0s in that dimension). This leaves 314 dimensions along which participants were clustered using the k-means algorithm (Lloyd, 1982). Figure 1 shows the distortion values for different numbers of clusters, revealing 3 clusters to be the inflection point at which more clusters provide only marginal returns. The algorithm takes this point as the true number of clusters because increasing the number of clusters beyond this simply subdivides the true clusters, thus over-fitting.

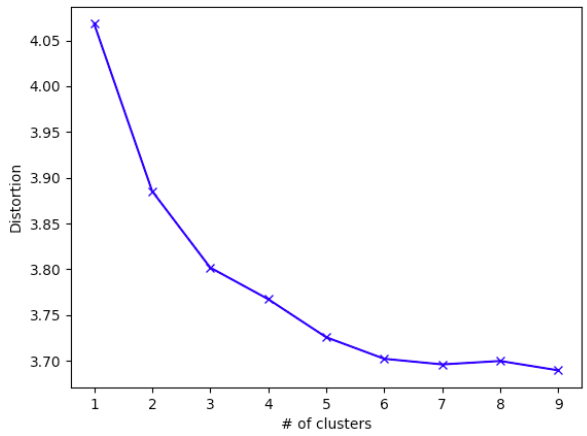

Figure 1: K-means clustering reveals 3 clusters of participants in our media consumption space. This is evidenced by the inflection in distortion decrease that occurs at $k = 3$.

Corpora Construction

We aggregate language data from the sources participants reported in our survey for further linguistic analysis. This produces two corpora (one for each cluster) that allow us to model their language differences. We fully acknowledge the difference between consuming sources as text, as our models do, and speech, as our participants do. Despite this, however, text fully captures the regularities of lexical and supra-lexical features we expect to influence performance on our behavioral tasks.

We collected each corpus by scraping repositories of television scripts (*Springfield! Springfield!*), movie subtitles (*YIFY Subtitles*), and song lyrics (the *Genius API*). For TV and music, we collected all the content for one show (e.g. all the scripts from *Law & Order*) or one artist (e.g. all the songs by *Bruno Mars*). We then pre-process these sources by removing anything the viewer would not hear (e.g. stage directions) and anything non-linguistic (e.g. non-alphanumeric characters or non-verbal noises).

Language and Surprisal Modeling

To model the language statistics of each cluster's corpus, we use 5-gram language models with backoff (Katz, 1987). These models estimate the likelihood of a sentence as the product of the conditional probabilities of its words given the words that precede them. Thus for a sentence of length L, the likelihood is:

$$\prod_{l=1}^{L} P(w_l | w_{l-(n-1)} \ldots w_{l-1}) \tag{1}$$

where n is a hyperparameter set to control the number of preceding words considered for context ($n = 1$ is simply the marginal probability). Because the probability of encountering the preceding string of words in training decreases as the length of the string increases, backoff allows the algorithm to decrease n until the preceding string *has* been seen in training (thus allowing the conditional probability to be estimated). Therefore, while we initially set our $n = 5$, probabilities may be calculated given less prior context.

In addition to the 5-gram model which proceeds from the beginning of a sentence seeking to model its probability, we also model the surprisal associated with encountering the final word of the sentence. This is a particularly important quantity considering both our behavioral tasks use sentence final words as their testing target. While in theory the model aligns with the concept of cloze probability – the probability of the sentence-final word given every preceding word: $P(w_L | w_1 \ldots w_{L-1})$ – this rarely occurs in practice given the sparseness of a training corpus. To model this, we adopt a similar method to n-gram models with backoff. We calculate the conditional probability of the last word given the $n-1$ preceding terms:

$$P(w_L | w_{L-(n-1)} \ldots w_{L-1}) \tag{2}$$

where we initialize $n = 5$ and reduce its value until the preceding string has been encountered in the training corpus ($n = 1$ is simply the marginal probability of the word occurring sentence-finally).

Results

Clustering

The clustering included all reported media sources and revealed three clusters based on participants' consumption habits. Despite a substantial drop in distortion from 2 clusters to 3 (see Fig. 1 for distortions), cluster 0 proved too small to analyze: it contains just 2 participants. Its size precludes both behavorial analysis, which requires an adequate number of samples to be statistically feasible, and computational modeling, which requires a corpus built from an adequate number of reported sources (aggregated across a cluster). Given these limitations, the following analyses will only use clusters 1 and 2 as the sample population (still 98% of the original sample). This clustering, far from an artifact of random seed, proved stable across random restarts. Over 1000 iterations, on average 75% of participants were re-clustered in the same groups (see **Behavioral Data** for the effects on statistical tests).

Regarding cluster membership, we expected USC students and community members to be unevenly distributed between clusters, and this was true, although not categorically. As seen in Table 1, the two are relatively balanced across clusters. Thus, cluster membership and *a priori* group membership are treated as orthogonal in the following analyses.

In addition to the a priori population, we examined the distribution of traditionally considered covariates across the clusters. We wanted to test whether self-reported media consumption provided new information beyond existing measures (i.e. we were not just capturing an existing highly correlated dimension of variance). As seen in Table 1, typical demographic variables were fairly evenly distributed across the clusters. One-way chi-square tests revealed that none of the demographic variables significantly differed from an even

split across clusters (i.e. the expected values if cluster and variable were independent).

Variable	Level	Cluster Ns		Cluster %	
		1	2	1	2
Population	USC	34	24	59%	41%
	LATTC	7	13	35%	65%
Gender[1]	Female	33	21	61%	39%
	Male	7	16	30%	70%
Schooling	High School	10	10	50%	50%
	Associate	4	4	50%	50%
	Some College	19	15	56%	44%
	Bachelor's	7	6	54%	46%
	Master's	1	2	33%	66%
Mono-lingual	True	10	11	48%	52%
	False	31	26	54%	46%
SES Self-Report[2]	High	13	13	50%	50%
	Medium	16	10	62%	38%
	Low	12	14	46%	54%

Table 1: The distribution of traditionally considered covariates across clusters is fairly even. We observe no obvious imbalance between clusters along any demographic dimensions our survey measured. One-way chi-square tests support this.

Given the orthogonality of self-reported media consumption to traditional demographic variables, we hereafter focus on the observed dimension of variance: media diet. We probe how the clusters differ in their media habits in order to delineate their makeup. We examine the clusters' centroids to calculate which dimensions (i.e. sources) they differ maximally along. This provides a measure of which media sources are most distinct between clusters. We find the following sources to be the 5 most different between clusters 1 and 2 and provide the difference in mean consumption between the two (i.e. $\bar{x}_1 - \bar{x}_2$) in parentheses: Star Wars (.64, specific films reported in the series were less powerful, on the order of .11-.17), Yes! (-.47), CNN (-.26), People (-.12), and Harry Potter (-.12). We hesitate to draw any conclusive generalities on the two clusters' media diets, but at a glance it appears that cluster 1 consumes lots of high fantasy (Star Wars, Lord of the Rings, The Chronicles of Narnia, etc.) while cluster 2 consumes more nonfiction (Yes!, CNN, People, etc.).

Behavioral Data

As seen in Figure 2, the SPiN task revealed a main effect of cluster, $F(1,76) = 7.30$, $p < .01$, but no main effect of population and no interaction between the two. In the reading span data, we again find a main effect of cluster, $F(1,76) = 4.05$, $p < .05$, and a main effect of population $F(1,76) = 13.57$, $p < .001$, and no interaction between the two. In our 1000 clustering iterations, 63% of iterations revealed statistically significant effects of cluster on the SPiN task (at $\alpha = .05$). This was not replicated with the span task, however: only 4% of our iterations found statistically significant effects.

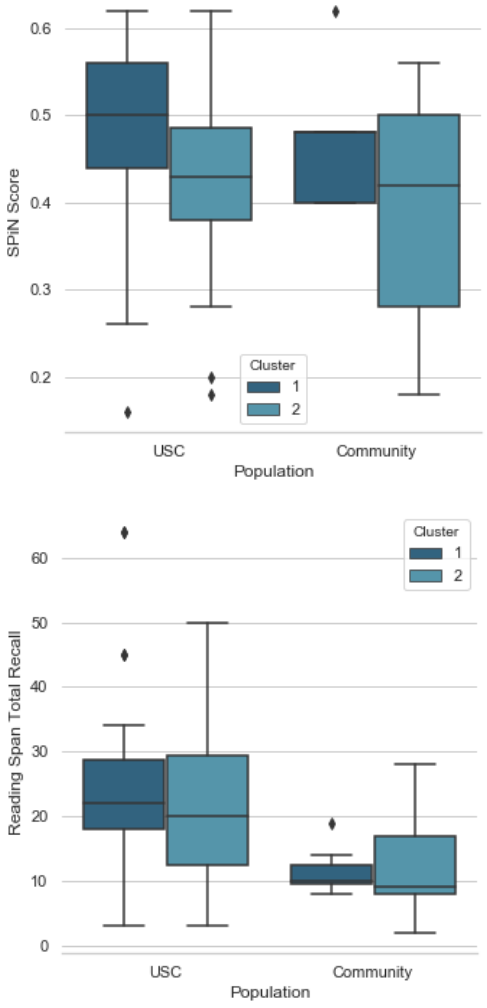

Figure 2: Results from the SPiN test reveal a significant difference between clusters but not populations. Reading span also shows an effect of cluster, but a larger effect of population. We observe no significant interactions.

The span test will play a minor role in further analyses, due to the difficulty in handling test result data and its scoring. Because the span task is terminated whenever participants fail to recall a set, participants provide unequal numbers of observations. The analyses are additionally constrained by the small number of items a typical participant completes. While observations exist for items later in the test, they are for a few extraordinary participants. This presents a problem not only in the paucity of observations, but also in the fact that these participants are unrepresentative of the general sample in their task abilities. As such, both item-level statistics and graphical representations are challenging.

Our survey obtains several pieces of demographic infor-

[1] One participant in cluster 1 chose not to report gender.

[2] Participants reported their SES on a continuous scale. Here, we bin responses into 3 quantiles to report distribution across clusters.

mation that are traditionally considered relevant covariates of performance on our cognitive tasks, such as socioeconomic status (SES, self-reported), age, education level, and monolingual status. None of these correlated significantly with performance on either task.

Language Media Input Modality

The above findings of differences between cluster performances motivated us to explore differences between clusters' survey behavior (other than the categorical responses which were used in clustering) to explain their performance data. In particular, we wondered whether the stronger task performances of cluster 1 might be due to increased experience with the tasks of speech perception and reading.

To probe this, we tested whether cluster 1 reported significantly more speech sources (TV, Movies, Music, and News shows) and significantly more text sources (Books, Newspapers, Magazines, Online News, and other online reading) than cluster 2. Indeed, we find that cluster 1 participants report significantly more listening on average than cluster 2: $t(42.82) = 3.09, p =< .005, d = 0.67$ (a medium effect). We also find that cluster 1 participants report significantly more reading on average than cluster 2: $t(72.6) = 5.10, p < .001, d = 1.13$ (a large effect). This may indicate an effect of modality-specific training on task performance. To probe this, we test the correlation between the number of speech sources a participant reports and their SPiN task performance. We test rank correlation rather than linear correlation as we are unsure of the linearity of the relationship between number of sources and modality-specific benefit, as well as to control for the effect of outliers in both performance and reporting volume. We observe a significant correlation between the two: $\rho(76) = .31, p = .005$. We do not, however, observe a significant correlation between number of text sources and span performance.

We also tested whether past modality preference (solicited with "when you were growing up...") would relate to current modality preference. We find a strong correlation between the amount of spoken language items reported growing up and amount of current items reported: $r(76) = .91, p < .001$. This correlation extends to the number of written language items, although not as strongly: $r(76) = .49, p < .001$.

Language Models

To evaluate the claim that our language models were capturing meaningful statistical regularities in the language of each cluster's corpus, we tested whether the log-likelihood produced by a model for each of the test items would correlate with mean performance on those items for the cluster. We do not observe a significant correlation between cluster 1's 5-gram model and performance on either the SPiN ($r(48) < 0.01$) or span ($r(25) = -0.01$). We also observe no significant correlation between cluster 2's 5-gram model and its performances on SPiN ($r(48) = -0.02$) or span ($r(25) < 0.01$). Additionally, we tested the correlation between cluster 1's 5th-order surprisal model and its performance and found no correlations with SPiN ($r(48) = 0.07$) or span ($r(25) = -0.09$). Similar results were obtained for cluster 2 (SPiN: $r(48) = 0.21$, span: $r(25) = 0.02$).

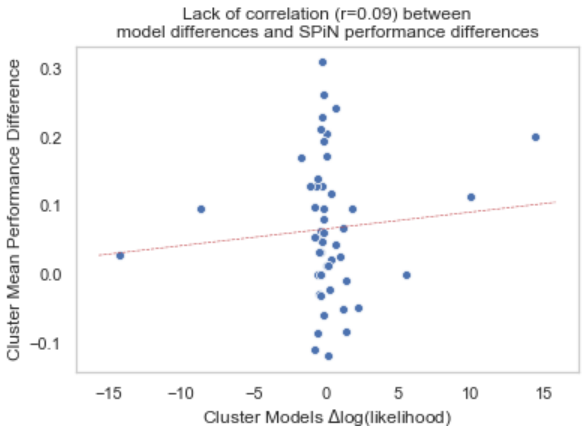

Figure 3: The non-correlation of cluster performance differences with model likelihood differences indicates the statistical information captured by the models is a poor predictor of behavioral performance. The significant cluster performance difference can be seen here by the majority of items occurring above 0-difference on the y-axis. A LMS-Regression line is drawn in red for reference.

In addition to modeling statistical properties of particular items, we also tested whether the difference between the language and surprisal models might capture the significant differences we see on our behavioral tasks. This method avoids any idiosyncrasies of particular items (as comparisons are within item) and instead captures any language differences of media sources. We again find a lack of significant correlation between 5-gram likelihood differences and task performance for both the SPiN ($r(48) = 0.09$) and span ($r(25) = .25$). Similar results are observed for the 5th-order surprisal model (SPiN: $r(48) < 0.01$, span: $r(25) = 0.08$). As shown in Fig. 3, differences between the model likelihoods are close to zero for most items, with a few outliers.

To examine the non-correlations and clustering around 0 on Fig. 3's x-axis, we tested the correlation between models and found strong correlations for both the SPiN ($r(48) = 0.94, p < .001$) and span ($r(86) = 0.97, p < .001$) test. These strong correlations, coupled with linear regression slopes of $\beta_1 = 0.96$ (SPiN) and $\beta_1 = 0.94$ (span) imply nearly identical log-likelihood scores between models despite training on categorically different sources. While the results reported here are specific to 5-gram language models and 5th-order surprisal models, other lower-ordered models of both yielded similar results.

Discussion

We observed significant performance differences on a speech perception in noise task and a working memory task between

clusters of participants derived from self-reported media consumption. These differences were above and beyond differences driven by *a priori* participant groups – students at a university vs. participants from the surrounding community. This clustering was robust to randomness and orthogonal to any traditionally considered demographic variables. As we have no reason to believe that the tests' target constructs systematically vary between our clusters, we conclude that media diet represents an uncorrelated latent variable moderating task performance. To our knowledge, our identifying media consumption as a significant orthogonal predictor of cognitive task performance is a novel contribution of this work.

This novel predictor is surprisingly powerful at explaining language test performance considering its complete lack of explicit linguistic information. In pursuing a linguistic explanation for our finding, we used statistical language models trained on sources participants reported consuming to analyze test items. These models did not identify particular stimuli driving performance differences, and we found no obvious differences in how well stimuli fit our models. However, a follow-up study we performed with more complex recurrent neural models did in fact reveal a correlation between models trained on our media corpora and behavioral performance (Courtland et al., 2019). This implies the statistics used here are not sophisticated enough. Cloze probability, for example, is computed as a simple ratio of the tokens of a word in context to all tokens in that exact context in the corpus.

Also of note is the highly significant difference in the number of sources reported by cluster 1 compared with cluster 2. It is possible that the greater number of sources indicates that cluster 1 contains more voracious consumers of media than cluster 2. This increased media consumption in the modalities of our tests may be providing cluster 1 members with modality-specific training they leverage at test time. Indeed, the correlation we observed between number of speech sources reported and SPiN performance supports this explanation. This is especially plausible given that watching a TV show or movie involves perceiving character dialog often obscured by various sources of noise (soundtrack, sound effects, etc.). It is also possible, however, that the increased responses and performance from cluster 1 is indicative not of their increased modality-specific training but rather a latent variable such as attentiveness or enthusiasm at participating in all aspects of the study.

It should be mentioned that participants' responses may reflect a (possibly implicit) choice to make specific habits known in the context of the survey. Given the importance of shared experience in forming relationships, what pieces and types of information people share and what they keep private often acts as a type of signaling that forms the basis of social cohesion. Thus, media diet survey responses may be more appropriately interpreted as signalling membership in a language community than literally reflecting the language practices of that community. Indeed, the vast majority of items in the corpora are professionally produced texts, which are likely to differ less than spontaneous spoken and written communication. In future work, we plan to obtain rich, naturalistic language samples in addition to the media corpora included so far to strengthen the evidence found here.

The identification of a dimension (other than the target construct) that test performance differs significantly along brings into question not only specific test validity probed here, but also the validity of the entire practice of test item standardization. This is true whether this dimension is categorical media consumption, shown here, or the linguistic content of the media, shown in Courtland et al. (2019). Tests that use language to probe target constructs must take the language of their test into account – not as a static entity to be standardized, but as the diverse and dynamic communication medium that it is. Test validity relies on the ability to generalize a test's result to participants' everyday behavior. This is only valid if the test is representative of the language they encounter in their daily lives (Coleman, 1964). Thus, tests employing standardized language not only contain inherent inequity for those less familiar with the test language, they are also less valid.

Here we aim to show that participants' diverse language experiences must be taken into account when diagnostic tools like those tested here are designed. Ideally, given the unique nature of language experience, test creators should strive to create tests that present equal difficulty to each participant by using personalized test language. This step to ensure equity is especially important given that test scores cannot simply be adjusted for using traditionally defined dialectal boundaries – as demonstrated here by the uninformativeness of the demographic variables that define these boundaries.

Generating equitable stimuli is a difficult or possibly infeasible task for human researchers, but could potentially be automated using generative models. If such models were driven by statistics that are highly representative of participants' language experience, they may do a better job of capturing cognitive constructs without smuggling in variability resulting from differences in language experience. Perhaps the most exciting future direction of this research will be to facilitate using more representative language statistics in designing stimuli for cognitive tests. The study of how language experience influences test performances that we take here represents a first step to understanding and mitigating this test inequity.

Acknowledgments

This work was supported by the NIH (5R21DC017018-02) and data were collected under USC IRB #UP-18-00006.

References

Akeroyd, M. A. (2008, January). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(sup2), S53–S71. doi: 10.1080/14992020802301142

American Speech-Language-Hearing Association (ASHA). (n.d.). *Cultural competence.*

Byrne, M. D. (1998, June). Taking a computational approach to aging: The SPAN theory of working memory. *Psychology and Aging*, *13*(2), 309–322. doi: 10.1037/0882-7974.13.2.309

Caspari, I., Parkinson, S. R., LaPointe, L. L., & Katz, R. C. (1998, July). Working Memory and Aphasia. *Brain and Cognition*, *37*(2), 205–223. doi: 10.1006/brcg.1997.0970

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school english in african american children and its relation to early reading achievement. *Child development*, *75*(5), 1340–1356.

Clopper, C. G., & Bradlow, A. R. (2008, September). Perception of Dialect Variation in Noise: Intelligibility and Classification. *Language and Speech*, *51*(3), 175–198. doi: 10.1177/0023830908098539

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*, 633–644.

Coleman, E. B. (1964, February). Generalizing to a Language Population. *Psychological Reports*, *14*(1), 219–226. doi: 10.2466/pr0.1964.14.1.219

Courtland, M., Davani, A., Reyes, M., Yeh, L., Leung, J., Kennedy, B., Dehghani, M., & Zevin, J. (2019). Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption. In *Proceedings of NLP+CSS: Workshops on Natural Language Processing and Computational Social Science,*. Minneapolis, MN.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466.

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017, December). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*. doi: 10.1177/2331216517744675

Elman, J. L. (Ed.). (2001). *Rethinking innateness: a connectionist perspective on development* (1. MIT Press paperback ed., 5. print ed.). Cambridge, Mass.: MIT Press.

Facts and Figures | About USC. (n.d.). Retrieved 2018-10-22, from `https://about.usc.edu/facts/`

Federmeier, K. D., Mai, H., & Kutas, M. (2005, July). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition*, *33*(5), 871–886. doi: 10.3758/BF03193082

Joseph, J. E., & Newmeyer, F. J. (2012). 'All languages are equally complex': The rise and fall of a consensus. *Historiographia Linguistica*, *39*(2-3), 341–368. doi: 10.1075/hl.39.2-3.08jos

Kalikow, D., Stevens, K., & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing* (pp. 400–401).

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., & MacDonald, M. C. (1998, October). Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. *Brain and Language*, *64*(3), 297–316. doi: 10.1006/brln.1998.1980

Kutas, M., & Hillyard, S. A. (1980, January). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. doi: 10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984, January). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter.

Lloyd, S. (1982, March). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. doi: 10.1109/TIT.1982.1056489

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior research methods*, *47*(4), 1095–1109.

Pellegrino, F., Coup, C., & Marsico, E. (2011). Across-Language Perspective on Speech Information Rate. *Language*, *87*(3), 539–558. doi: 10.1353/lan.2011.0057

Schneider, E. W., & Kortmann, B. (Eds.). (2004). *A handbook of varieties of English: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter.

Seidenberg, M. S., & MacDonald, M. C. (1999, October). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, *23*(4), 569–588. doi: 10.1016/S0364-0213(99)00016-6

Stone, M., Gabrieli, J. D. E., Stebbins, G. T., & Sullivan, E. V. (1998, April). Working and strategic memory deficits in schizophrenia. *Neuropsychology*, *12*(2), 278–288. doi: 10.1037/0894-4105.12.2.278

US Census Bureau. (n.d.). *American Community Survey Data.* Retrieved 2018-10-22, from `https://www.census.gov/programs-surveys/acs/data.html`

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.